Computing similarity between text and human summaries. A comparison of models

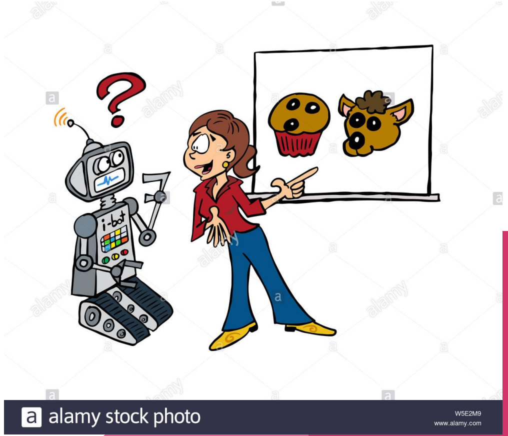
Marianela Crissman

Motivation

To increase education

- Online Learning
- Two-way communication => Scalability issue
- Intelligent Tutoring Systems

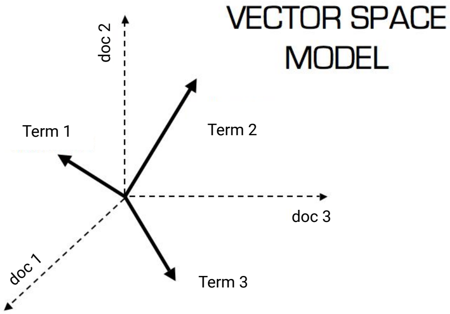
Comparing user-response vs text book-response.



Background

Vector Modeling

1. Corpus
    - d1 = “computer master tutor”
      d2 = “computer system learning response”
      d3 = ”learning tutor system”
      :
2. Bag of words: unique tokens
3. Term x Document matrix



	d1	d2	d3	d4	d5	d6	d7
master	1	0	0	1	0	0	0
tutor	1	0	1	0	0	0	0
computer	1	1	0	0	0	0	0
learning	0	1	1	0	1	0	0
system	0	1	1	2	0	0	0
response	0	1	0	0	1	0	0

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Similarity computed using Cosine between vectors
High dimension sparse matrices, certain relationships are not captured.

Background

Latent Semantic Analysis (LSA)

- Words that appear together are somehow related. LSA captures this relationship and includes it into the vector model.

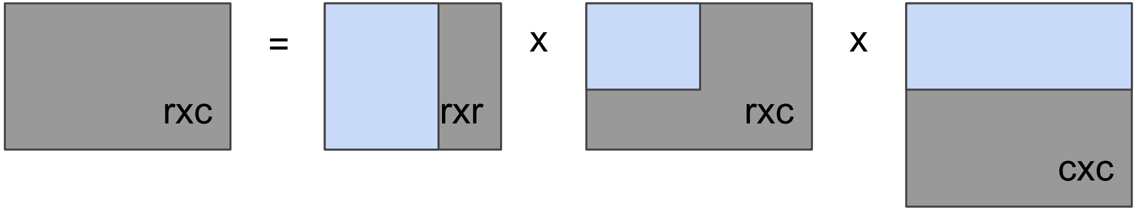
Work on Latent Semantic Analysis (LSA):

- Deerwester et al (1990). Landauer et al (1998) - Latent Semantic Analysis
- Steinberger et al (2012) - Evaluation Measure for Text Summarization
- Nye, B.D., Graesser (2014) - 17 years of AutoTutor

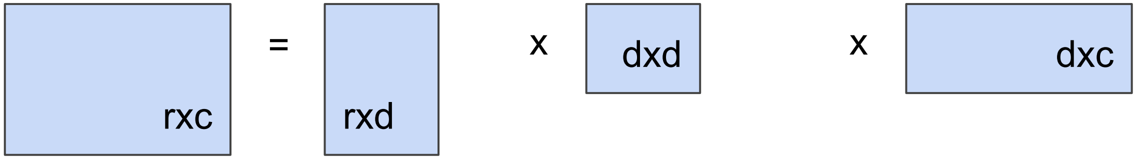
Background

LSA uses dimensionality reduction. Good enough: d=100..400.

Singular Value Decomposition (SVD)

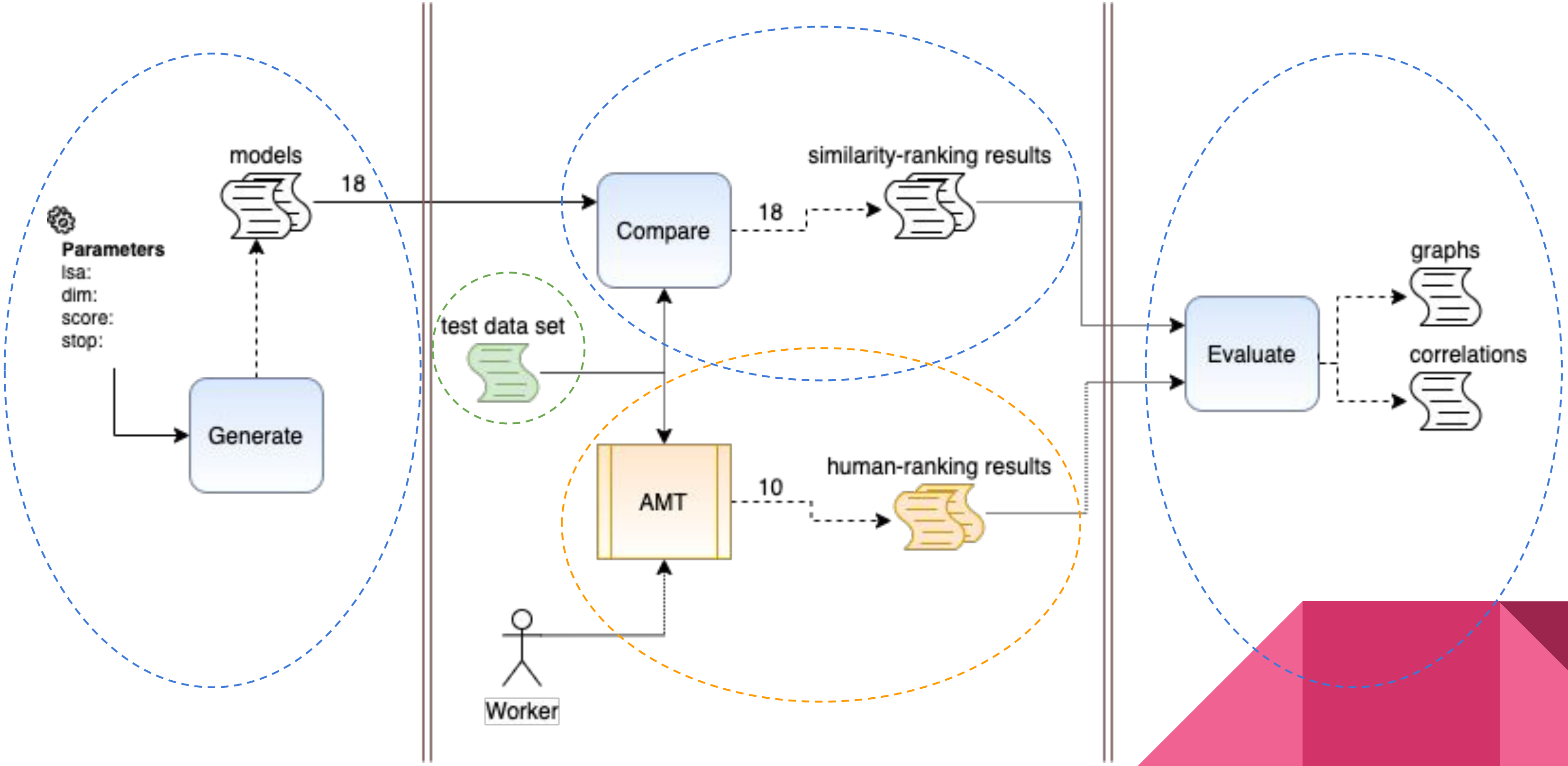


Reconstruction

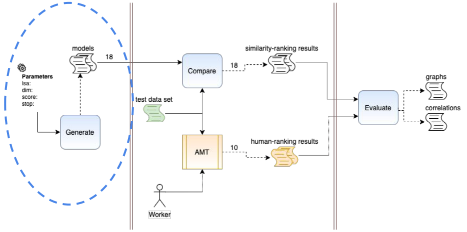


Words that often appeared together in documents would have vectors pointing to similar directions.

Methods


models
18
Parameters
lsa:
dim:
score:
stop:
Generate
Compare
18
similarity-ranking results
test data set
AMT
10
human-ranking results
Worker
Evaluate
graphs
correlations

Methods - Generate



1. Corpus: a collection of senate speeches. Data from (Diermeier et al. 2012)

2. Pre-process: Punctuation, Lowercase, @YEAR@ and @PRICE@

3. Vectorization:

   a. Bag of words: [a-zA-Z0-9]{2,}; unigrams; appeared in at least 2 documents

   b. Score: frequency, zero-one, TF-IDF

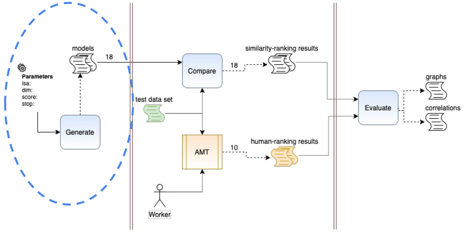
   c. Stop words: remove words that do not add meaning

4?. Apply LSA: d=100 or d=300 dimensions

	d1	d2	d3	d4	d5	d6	d7
master	1	0	0	1	0	0	0
tutor	1	0	1	0	0	0	0
computer	1	1	0	0	0	0	0
learning	0	1	1	0	1	0	0
system	0	1	1	2	0	0	0
response	0	1	0	0	1	0	0

Methods - Generate

VSM	dimensionality	score	stop words
1	-	count	yes
2	-	count	-
3	-	tfidf	yes
4	-	tfidf	-
5	-	zero-one	yes
6	-	zero-one	-
7	100	count	yes
8	100	count	-
9	100	tfidf	yes
10	100	tfidf	-
11	100	zero-one	yes
12	100	zero-one	-
13	300	count	yes
14	300	count	-
15	300	tfidf	yes
16	300	tfidf	-
17	300	zero-one	yes
18	300	zero-one	-



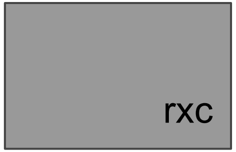
Vector Space Models

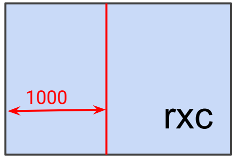
- lsa: used lsa or not
- dim: 100 or 300
- sco: count, tfidf, zero-one
- sw: used stop words or not

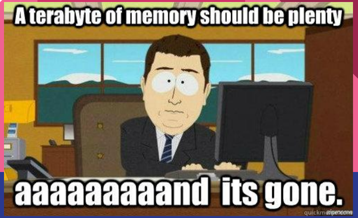
Size of a model: (60k, 130k) ~ 32.5GB

Storage:

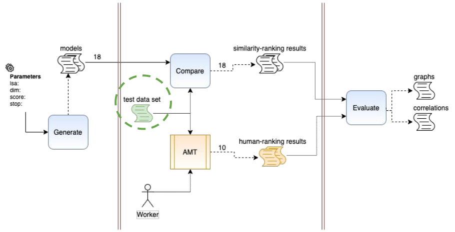
- Models 1-6 as sparse matrix.
- Models 7-18, kept first 1000 columns.







Methods - Test Data Set



From *Mi Guia**, we extracted 10 lead paragraphs (*LPx*). x=1.10
Six volunteers derived summaries (*S*) for each of the LPx.

- $S_{y,x}$ = Summary of volunteer y on LPx

To test
We grouped each LPx with 6 Summary Paragraphs (SP(i,x), i=1..6), where SP(i,x):

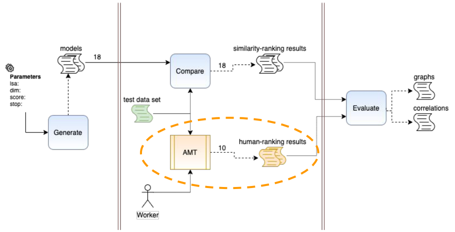
- was not similar for i=1,2
- was similar for i=3,4,5,6

Table 3.2.1. Example of a task to be ranked.

Paragraph: LP1	Cancer begins in cells. Cells are the building blocks that make up all parts of our bodies, including our breasts. Cancer begins when the cells in the breast change and grow into a tumor. If not removed or treated, a breast cancer tumor can spread to other parts of the body and become lethal. These days, there are many treatment options for breast cancer and the majority of women are cured of breast cancer.
sentence SP(1,1)	Systemic therapy reduces the chances of breast cancer coming back.
sentence SP(2,1)	Mastectomy is a procedure that eliminates the breast.
sentence SP(3,1)	Cancer starts in the smallest entity of our bodies; the cell grows too the parts of the body and if left untreated will probably end in death.
sentence SP(4,1)	Cancer is a pathologic function of cells, your body is made by cells. This means that you can have cancer in a specific part of your body or you can have cancer everywhere. When it starts, it is usually in a determined territory and it is easier to treat and get a better prognosis in that moment.
sentence SP(5,1)	Cancer starts in the body cells and when it grows out of control, it becomes a tumor. For breast cancer there are actually several treatment options that can prevent the spreading of malignant cells to other body organs, so a big number of women are cured from this disease.
sentence SP(6,1)	Sometimes cells in the body can change, and may grow into cancerous tumors, which must be removed from the body before they become lethal.

* An application for Breast Cancer education.

Methods - Human’s perception

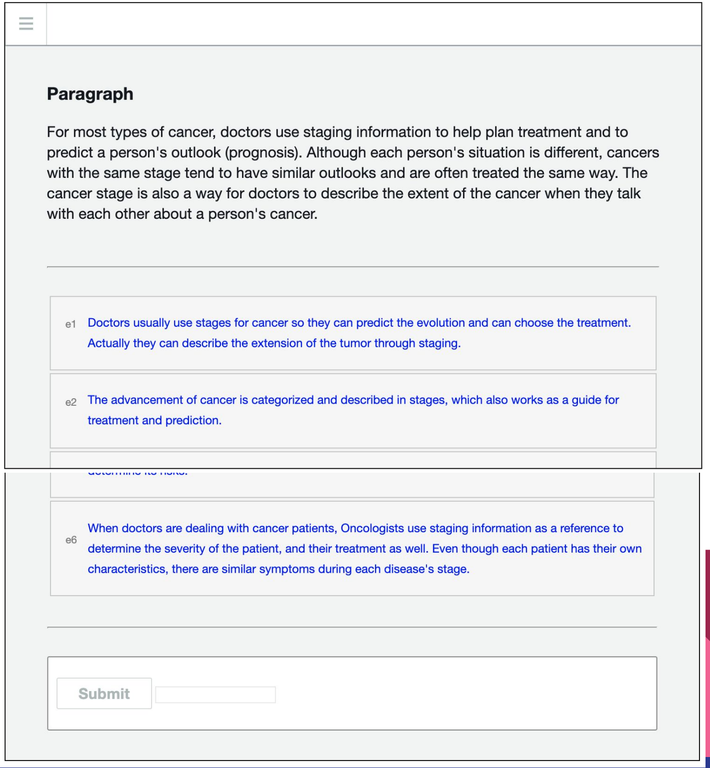


From each LPx, we created one Human Intelligence Tasks (HIT) in Amazon Mech. Turk

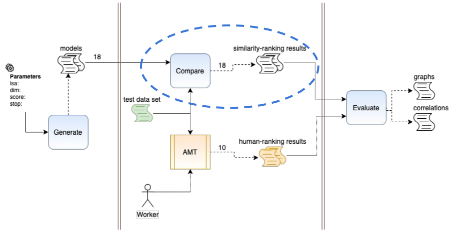
- Each HIT completed by 10 humans

Task: order the blocks depending on how similar they consider these sentences are to the paragraph.

They ended up ranking SP(1,x) - SP(6,x).



Methods - Automatic Similarity

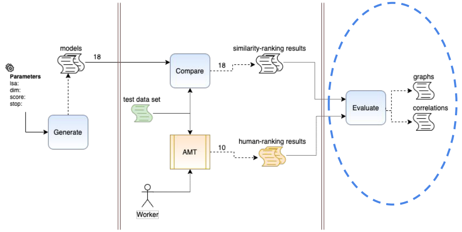


Given a paragraph, a summary paragraph and a model: **LPx**, **SP(i,x)** and **m**.

1. Preprocessed texts:
   - {?!, lower, @PRICE@, @YEAR@}
2. Vectorized the sentence, using model **m**
   - Sentence vector = mean(word vectors).
3. Compared LPx paragraph to each one of its SP(i,x),
   - Cosine Similarity(LPx, SP(i,x)) , for x=1..10 and i=1..6

Comparisons belonging to ***(LPx, m)*** were sorted and ranked:
1- most similar and 6- least similar.

Methods - Evaluation



Using model ***m****=1..18* and paragraph ***x****=1..10*.

Machine Rankings, $\boldsymbol{mr_{m,x}}$ = ***MRanking****(****m, LPx****)*. Human Rankings, $\boldsymbol{hr_x}$ = ***HRanking(LPx)***

For each ranking pair $(mr_{m,x}, hr_x)$:

- Spearman's Rank coefficient

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \qquad d_i = \mathrm{rg}(X_i) - \mathrm{rg}(Y_i)$$

- Kendall Tau Correlation

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j).$$

- Order Preservation Measure

$$orderPreservationMeasure = \frac{1}{C_{n,2}} \sum_{i=0}^{n-1} \sum_{j=i+1}^{n} isAscending(i,j,A_{num}),$$

$$isAscending(i,j,A_{num}) = 1 \quad if\ A_{num}[i] < A_{num}[j], \quad 0\ otherwise.$$

RESULTS

Results - Models

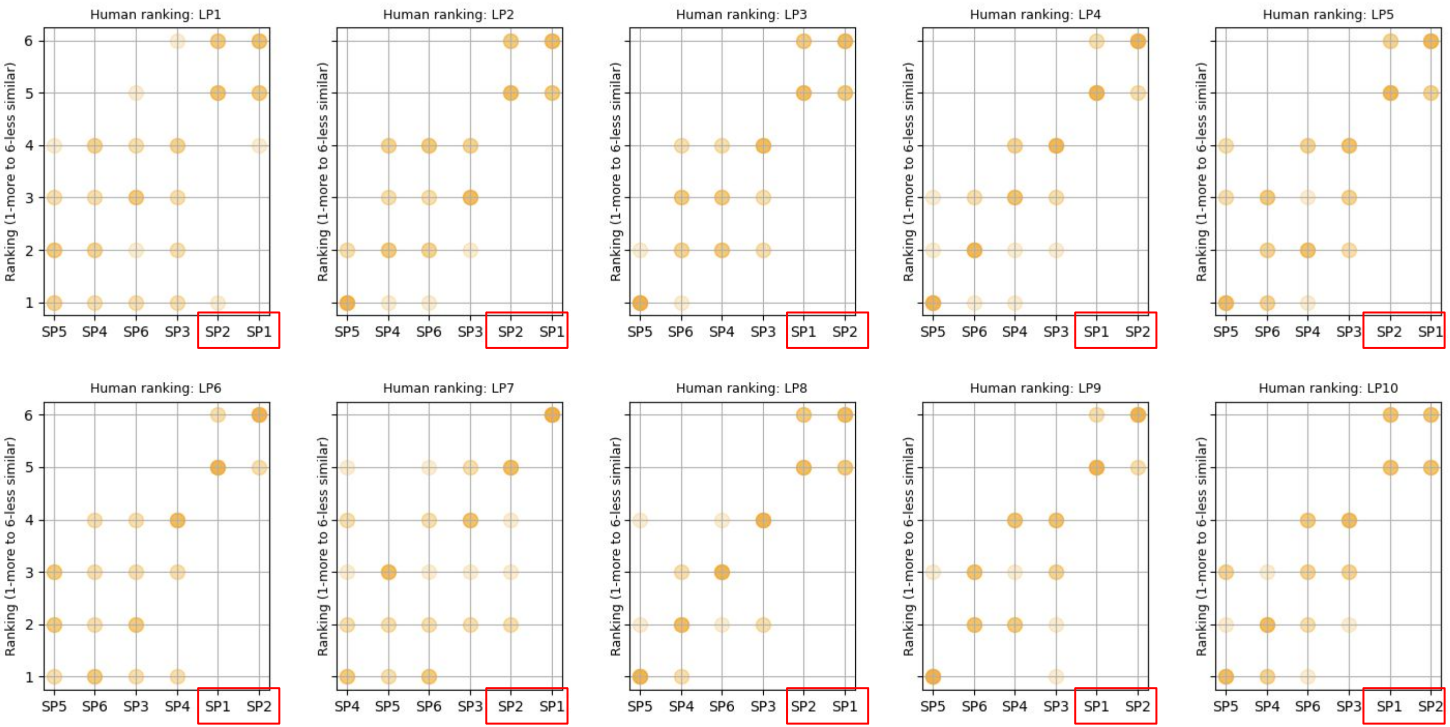
model	
lsa:- dim:- sco:count sw:yes	1
lsa:- dim:- sco:count sw:-	2
lsa:- dim:- sco:tfidf sw:yes	3
lsa:- dim:- sco:tfidf sw:-	4
lsa:- dim:- sco:zeroone sw:yes	5
lsa:- dim:- sco:zeroone sw:-	6
lsa:yes dim:100 sco:count sw:yes	7
lsa:yes dim:100 sco:count sw:-	8
lsa:yes dim:100 sco:tfidf sw:yes	9
lsa:yes dim:100 sco:tfidf sw:-	10
lsa:yes dim:100 sco:zeroone sw:yes	11
lsa:yes dim:100 sco:zeroone sw:-	12
lsa:yes dim:300 sco:count sw:yes	13
lsa:yes dim:300 sco:count sw:-	14
lsa:yes dim:300 sco:tfidf sw:yes	15
lsa:yes dim:300 sco:tfidf sw:-	16
lsa:yes dim:300 sco:zeroone sw:yes	17
lsa:yes dim:300 sco:zeroone sw:-	18

Generated 18 models

Each Vector Space Model:

- In approximately 5 minutes
- Each model took approximately 300 MB on disk
- Each model contained around 60k words.

Results - Human Rankings



Humans ranking of SP(i,x) per LPx.

- Agreed on SP1, SP2 less similar
- Confusion with SP3-SP6.

Opacity: 1- more votes, 0- less votes.

Results - Similarities

1 0.5 0

Cell: similarity between ***LPx*** and ***SP(i,x)***, using model ***m***.

model		LP1						LP2					
		SP5	SP4	SP3	SP6	SP1	SP2	SP5	SP6	SP4	SP1	SP2	SP3
lsa:- dim:- sco:count sw:yes	1	0.804	0.666	0.662	0.620	0.517	0.378	0.754	0.749	0.579	0.502	0.449	0.299
lsa:- dim:- sco:count sw:-	2	0.998	0.996	0.990	0.992	0.980	0.991	0.995	0.993	0.991	0.982	0.995	0.986
lsa:- dim:- sco:tfidf sw:yes	3	0.656	0.450	0.510	0.480	0.419	0.306	0.678	0.664	0.524	0.393	0.401	0.366
lsa:- dim:- sco:tfidf sw:-	4	0.993	0.988	0.975	0.975	0.956	0.975	0.986	0.969	0.972	0.953	0.982	0.962
lsa:- dim:- sco:zeroone sw:yes	5	0.750	0.658	0.695	0.686	0.428	0.253	0.854	0.837	0.722	0.553	0.535	0.392
lsa:- dim:- sco:zeroone sw:-	6	0.992	0.989	0.977	0.977	0.954	0.964	0.991	0.977	0.975	0.968	0.973	0.968
lsa:yes dim:100 sco:count sw:yes	7	0.926	0.868	0.786	0.826	0.615	0.407	0.711	0.747	0.491	0.635	0.552	0.244
lsa:yes dim:100 sco:count sw:-	8	0.998	0.996	0.990	0.991	0.979	0.989	0.994	0.991	0.988	0.979	0.995	0.986
lsa:yes dim:100 sco:tfidf sw:yes	9	0.829	0.830	0.738	0.725	0.525	0.444	0.729	0.760	0.520	0.635	0.584	0.299
lsa:yes dim:100 sco:tfidf sw:-	10	0.996	0.991	0.981	0.984	0.962	0.979	0.988	0.979	0.976	0.963	0.987	0.970
lsa:yes dim:100 sco:zeroone sw:yes	11	0.923	0.924	0.918	0.881	0.824	0.700	0.899	0.897	0.768	0.846	0.795	0.720
lsa:yes dim:100 sco:zeroone sw:-	12	0.997	0.994	0.992	0.986	0.976	0.974	0.996	0.988	0.982	0.980	0.980	0.979
lsa:yes dim:300 sco:count sw:yes	13	0.854	0.642	0.698	0.707	0.490	0.362	0.681	0.714	0.483	0.608	0.562	0.288
lsa:yes dim:300 sco:count sw:-	14	0.998	0.995	0.990	0.991	0.979	0.989	0.994	0.991	0.988	0.978	0.995	0.985
lsa:yes dim:300 sco:tfidf sw:yes	15	0.802	0.800	0.627	0.701	0.473	0.308	0.673	0.631	0.452	0.580	0.548	0.332
lsa:yes dim:300 sco:tfidf sw:-	16	0.995	0.990	0.980	0.981	0.962	0.978	0.987	0.974	0.974	0.959	0.985	0.968
lsa:yes dim:300 sco:zeroone sw:yes	17	0.832	0.725	0.773	0.742	0.692	0.562	0.862	0.829	0.707	0.713	0.647	0.689
lsa:yes dim:300 sco:zeroone sw:-	18	0.995	0.989	0.986	0.981	0.966	0.971	0.993	0.982	0.979	0.973	0.977	0.973

model		LP3						LP4					
		SP5	SP4	SP3	SP6	SP2	SP1	SP6	SP5	SP4	SP3	SP2	SP1
lsa:- dim:- sco:count sw:yes	1	0.911	0.707	0.539	0.506	0.494	0.369	0.821	0.818	0.739	0.661	0.611	0.427
lsa:- dim:- sco:count sw:-	2	0.996	0.993	0.988	0.991	0.982	0.986	0.985	0.996	0.992	0.996	0.994	0.995
lsa:- dim:- sco:tfidf sw:yes	3	0.770	0.639	0.471	0.455	0.440	0.368	0.693	0.682	0.657	0.572	0.425	0.322
lsa:- dim:- sco:tfidf sw:-	4	0.984	0.977	0.972	0.972	0.964	0.950	0.955	0.986	0.975	0.984	0.979	0.979
lsa:- dim:- sco:zeroone sw:yes	5	0.884	0.837	0.608	0.557	0.584	0.438	0.768	0.827	0.788	0.609	0.602	0.537
lsa:- dim:- sco:zeroone sw:-	6	0.986	0.981	0.969	0.978	0.971	0.959	0.967	0.988	0.983	0.979	0.978	0.975
lsa:yes dim:100 sco:count sw:yes	7	0.947	0.694	0.739	0.714	0.558	0.346	0.934	0.921	0.856	0.818	0.710	0.458
lsa:yes dim:100 sco:count sw:-	8	0.996	0.992	0.986	0.990	0.980	0.982	0.984	0.995	0.992	0.995	0.991	0.994
lsa:yes dim:100 sco:tfidf sw:yes	9	0.934	0.718	0.714	0.819	0.658	0.433	0.932	0.897	0.844	0.818	0.707	0.505
lsa:yes dim:100 sco:tfidf sw:-	10	0.993	0.983	0.977	0.984	0.969	0.965	0.973	0.990	0.981	0.987	0.984	0.984
lsa:yes dim:100 sco:zeroone sw:yes	11	0.964	0.894	0.819	0.904	0.868	0.803	0.916	0.926	0.879	0.877	0.759	0.895
lsa:yes dim:100 sco:zeroone sw:-	12	0.992	0.991	0.979	0.992	0.983	0.984	0.978	0.992	0.990	0.987	0.988	0.984
lsa:yes dim:300 sco:count sw:yes	13	0.928	0.678	0.585	0.630	0.492	0.260	0.884	0.874	0.808	0.749	0.649	0.430
lsa:yes dim:300 sco:count sw:-	14	0.996	0.992	0.986	0.990	0.979	0.982	0.981	0.995	0.992	0.995	0.991	0.994
lsa:yes dim:300 sco:tfidf sw:yes	15	0.898	0.671	0.614	0.676	0.453	0.240	0.883	0.824	0.751	0.780	0.560	0.449
lsa:yes dim:300 sco:tfidf sw:-	16	0.988	0.979	0.973	0.981	0.964	0.958	0.964	0.988	0.978	0.986	0.981	0.982
lsa:yes dim:300 sco:zeroone sw:yes	17	0.925	0.865	0.736	0.842	0.719	0.717	0.824	0.849	0.811	0.789	0.682	0.779
lsa:yes dim:300 sco:zeroone sw:-	18	0.987	0.985	0.972	0.983	0.976	0.974	0.969	0.988	0.985	0.982	0.981	0.978

model		LP5						LP6					
		SP5	SP4	SP2	SP6	SP3	SP1	SP5	SP4	SP2	SP1	SP6	SP3
lsa:- dim:- sco:count sw:yes	1	0.900	0.897	0.555	0.534	0.517	0.279	0.888	0.615	0.602	0.579	0.526	0.384
lsa:- dim:- sco:count sw:-	2	0.995	0.997	0.990	0.997	0.996	0.945	0.996	0.986	0.966	0.986	0.994	0.994
lsa:- dim:- sco:tfidf sw:yes	3	0.675	0.663	0.448	0.482	0.371	0.271	0.720	0.499	0.463	0.356	0.462	0.335
lsa:- dim:- sco:tfidf sw:-	4	0.987	0.988	0.972	0.990	0.988	0.837	0.987	0.969	0.905	0.962	0.977	0.985
lsa:- dim:- sco:zeroone sw:yes	5	0.809	0.810	0.666	0.697	0.546	0.325	0.861	0.716	0.709	0.645	0.706	0.497
lsa:- dim:- sco:zeroone sw:-	6	0.993	0.990	0.982	0.993	0.990	0.938	0.988	0.979	0.947	0.977	0.981	0.979
lsa:yes dim:100 sco:count sw:yes	7	0.967	0.958	0.645	0.515	0.708	0.282	0.945	0.690	0.722	0.569	0.691	0.521
lsa:yes dim:100 sco:count sw:-	8	0.993	0.995	0.986	0.997	0.995	0.928	0.996	0.986	0.959	0.984	0.992	0.993
lsa:yes dim:100 sco:tfidf sw:yes	9	0.949	0.927	0.677	0.567	0.666	0.344	0.900	0.768	0.775	0.535	0.739	0.569
lsa:yes dim:100 sco:tfidf sw:-	10	0.988	0.991	0.977	0.993	0.991	0.847	0.992	0.975	0.944	0.969	0.987	0.988
lsa:yes dim:100 sco:zeroone sw:yes	11	0.955	0.918	0.785	0.847	0.892	0.701	0.939	0.886	0.896	0.752	0.928	0.880
lsa:yes dim:100 sco:zeroone sw:-	12	0.997	0.996	0.988	0.997	0.994	0.985	0.996	0.988	0.969	0.986	0.993	0.985
lsa:yes dim:300 sco:count sw:yes	13	0.950	0.941	0.564	0.504	0.651	0.297	0.907	0.620	0.647	0.548	0.644	0.479
lsa:yes dim:300 sco:count sw:-	14	0.993	0.995	0.986	0.997	0.995	0.930	0.996	0.986	0.958	0.984	0.992	0.993
lsa:yes dim:300 sco:tfidf sw:yes	15	0.904	0.865	0.583	0.552	0.634	0.386	0.803	0.655	0.664	0.435	0.689	0.542
lsa:yes dim:300 sco:tfidf sw:-	16	0.988	0.990	0.974	0.993	0.990	0.837	0.990	0.973	0.914	0.963	0.984	0.987
lsa:yes dim:300 sco:zeroone sw:yes	17	0.875	0.876	0.689	0.745	0.582	0.620	0.900	0.831	0.785	0.710	0.845	0.785
lsa:yes dim:300 sco:zeroone sw:-	18	0.995	0.992	0.984	0.995	0.991	0.947	0.991	0.985	0.956	0.979	0.988	0.982

model		LP7						LP8					
		SP4	SP3	SP6	SP1	SP5	SP2	SP5	SP3	SP2	SP6	SP4	SP1
lsa:- dim:- sco:count sw:yes	1	0.656	0.652	0.630	0.548	0.538	0.498	0.891	0.657	0.612	0.591	0.409	0.246
lsa:- dim:- sco:count sw:-	2	0.996	0.997	0.996	0.981	0.997	0.992	0.990	0.996	0.993	0.996	0.995	0.975
lsa:- dim:- sco:tfidf sw:yes	3	0.612	0.574	0.668	0.462	0.551	0.488	0.827	0.634	0.526	0.533	0.410	0.130
lsa:- dim:- sco:tfidf sw:-	4	0.984	0.989	0.983	0.954	0.990	0.973	0.968	0.983	0.978	0.986	0.982	0.921
lsa:- dim:- sco:zeroone sw:yes	5	0.779	0.726	0.753	0.640	0.669	0.597	0.601	0.698	0.551	0.705	0.668	0.488
lsa:- dim:- sco:zeroone sw:-	6	0.983	0.979	0.981	0.937	0.982	0.981	0.971	0.978	0.979	0.979	0.977	0.953
lsa:yes dim:100 sco:count sw:yes	7	0.659	0.794	0.684	0.509	0.622	0.600	0.963	0.675	0.680	0.722	0.424	0.301
lsa:yes dim:100 sco:count sw:-	8	0.996	0.997	0.997	0.983	0.997	0.990	0.988	0.995	0.995	0.995	0.994	0.969
lsa:yes dim:100 sco:tfidf sw:yes	9	0.690	0.782	0.710	0.543	0.710	0.652	0.958	0.559	0.699	0.721	0.429	0.219
lsa:yes dim:100 sco:tfidf sw:-	10	0.991	0.994	0.991	0.963	0.994	0.982	0.971	0.987	0.985	0.989	0.988	0.943
lsa:yes dim:100 sco:zeroone sw:yes	11	0.867	0.793	0.858	0.702	0.878	0.868	0.882	0.795	0.912	0.906	0.854	0.674
lsa:yes dim:100 sco:zeroone sw:-	12	0.991	0.984	0.989	0.958	0.987	0.994	0.985	0.988	0.991	0.988	0.989	0.962
lsa:yes dim:300 sco:count sw:yes	13	0.580	0.648	0.542	0.515	0.622	0.606	0.900	0.690	0.562	0.704	0.458	0.344
lsa:yes dim:300 sco:count sw:-	14	0.996	0.997	0.996	0.982	0.997	0.990	0.988	0.995	0.994	0.995	0.994	0.968
lsa:yes dim:300 sco:tfidf sw:yes	15	0.607	0.622	0.648	0.477	0.697	0.659	0.911	0.555	0.548	0.677	0.488	0.265
lsa:yes dim:300 sco:tfidf sw:-	16	0.987	0.990	0.986	0.959	0.992	0.979	0.970	0.986	0.982	0.987	0.985	0.923
lsa:yes dim:300 sco:zeroone sw:yes	17	0.827	0.762	0.743	0.671	0.722	0.605	0.649	0.735	0.635	0.745	0.766	0.562
lsa:yes dim:300 sco:zeroone sw:-	18	0.985	0.979	0.983	0.947	0.983	0.984	0.974	0.982	0.982	0.982	0.979	0.954

model		LP9						LP10					
		SP5	SP4	SP3	SP2	SP6	SP1	SP6	SP3	SP5	SP4	SP2	SP1
lsa:- dim:- sco:count sw:yes	1	0.694	0.520	0.477	0.460	0.409	0.388	0.752	0.746	0.638	0.621	0.476	0.414
lsa:- dim:- sco:count sw:-	2	0.996	0.996	0.996	0.984	0.997	0.994	0.993	0.998	0.997	0.995	0.993	0.996
lsa:- dim:- sco:tfidf sw:yes	3	0.548	0.385	0.356	0.301	0.413	0.394	0.640	0.641	0.569	0.533	0.305	0.286
lsa:- dim:- sco:tfidf sw:-	4	0.985	0.980	0.985	0.955	0.990	0.977	0.980	0.991	0.993	0.984	0.979	0.986
lsa:- dim:- sco:zeroone sw:yes	5	0.665	0.468	0.414	0.452	0.288	0.302	0.746	0.782	0.729	0.698	0.453	0.401
lsa:- dim:- sco:zeroone sw:-	6	0.982	0.973	0.973	0.968	0.986	0.962	0.973	0.985	0.994	0.980	0.964	0.978
lsa:yes dim:100 sco:count sw:yes	7	0.920	0.806	0.850	0.688	0.410	0.379	0.889	0.910	0.751	0.782	0.775	0.717
lsa:yes dim:100 sco:count sw:-	8	0.995	0.995	0.995	0.981	0.997	0.993	0.992	0.998	0.996	0.995	0.993	0.996
lsa:yes dim:100 sco:tfidf sw:yes	9	0.916	0.828	0.816	0.649	0.455	0.432	0.892	0.902	0.738	0.761	0.716	0.675
lsa:yes dim:100 sco:tfidf sw:-	10	0.988	0.986	0.989	0.963	0.993	0.981	0.983	0.995	0.994	0.989	0.985	0.990
lsa:yes dim:100 sco:zeroone sw:yes	11	0.940	0.749	0.868	0.760	0.843	0.678	0.961	0.953	0.854	0.935	0.851	0.889
lsa:yes dim:100 sco:zeroone sw:-	12	0.992	0.985	0.989	0.985	0.994	0.975	0.988	0.991	0.998	0.994	0.991	0.989
lsa:yes dim:300 sco:count sw:yes	13	0.891	0.707	0.833	0.650	0.441	0.393	0.886	0.853	0.721	0.726	0.631	0.645
lsa:yes dim:300 sco:count sw:-	14	0.995	0.995	0.995	0.981	0.997	0.993	0.992	0.998	0.996	0.995	0.993	0.995
lsa:yes dim:300 sco:tfidf sw:yes	15	0.822	0.667	0.734	0.604	0.478	0.436	0.882	0.875	0.698	0.783	0.618	0.628
lsa:yes dim:300 sco:tfidf sw:-	16	0.987	0.982	0.986	0.956	0.992	0.980	0.982	0.993	0.994	0.987	0.984	0.989
lsa:yes dim:300 sco:zeroone sw:yes	17	0.798	0.489	0.484	0.489	0.553	0.457	0.906	0.845	0.775	0.851	0.517	0.672
lsa:yes dim:300 sco:zeroone sw:-	18	0.986	0.976	0.979	0.973	0.990	0.965	0.984	0.988	0.997	0.990	0.981	0.982

On each ***LPx*** per model ***m***, we are looking for:

- Two not related.
- Four related.

Results - Similarities

Similarities on LP1. SP(i,1) i=1..6.

- Alternation:
    - Stop words remove noise.
- Model's threshold:
    - Model 11: values 0.70 - 0.92
    - Model 3: values 0.30 - 0.65

Each model has a different threshold for determining when something is similar or not.

1 0.5 0

model		LP1					
		SP5	SP4	SP3	SP6	SP1	SP2
lsa:- dim:- sco:count sw:yes	1	0.804	0.666	0.662	0.620	0.517	0.378
lsa:- dim:- sco:count sw:-	2	0.998	0.996	0.990	0.992	0.980	0.991
lsa:- dim:- sco:tfidf sw:yes	3	0.656	0.450	0.510	0.480	0.419	0.306
lsa:- dim:- sco:tfidf sw:-	4	0.993	0.988	0.975	0.975	0.956	0.975
lsa:- dim:- sco:zeroone sw:yes	5	0.750	0.658	0.695	0.686	0.428	0.253
lsa:- dim:- sco:zeroone sw:-	6	0.992	0.989	0.977	0.977	0.954	0.964
lsa:yes dim:100 sco:count sw:yes	7	0.926	0.868	0.786	0.826	0.615	0.407
lsa:yes dim:100 sco:count sw:-	8	0.998	0.996	0.990	0.991	0.979	0.989
lsa:yes dim:100 sco:tfidf sw:yes	9	0.829	0.830	0.738	0.725	0.525	0.444
lsa:yes dim:100 sco:tfidf sw:-	10	0.996	0.991	0.981	0.984	0.962	0.979
lsa:yes dim:100 sco:zeroone sw:yes	11	0.923	0.924	0.918	0.881	0.824	0.700
lsa:yes dim:100 sco:zeroone sw:-	12	0.997	0.994	0.992	0.986	0.976	0.974
lsa:yes dim:300 sco:count sw:yes	13	0.854	0.642	0.698	0.707	0.490	0.362
lsa:yes dim:300 sco:count sw:-	14	0.998	0.995	0.990	0.991	0.979	0.989
lsa:yes dim:300 sco:tfidf sw:yes	15	0.802	0.800	0.627	0.701	0.473	0.308
lsa:yes dim:300 sco:tfidf sw:-	16	0.995	0.990	0.980	0.981	0.962	0.978
lsa:yes dim:300 sco:zeroone sw:yes	17	0.832	0.725	0.773	0.742	0.692	0.562
lsa:yes dim:300 sco:zeroone sw:-	18	0.995	0.989	0.986	0.981	0.966	0.971

Results - Evaluating Rankings

Spearman's Correlation

Blue: *1*(dark) to *-1*(white)

Computed average and standard deviation of results, per model.

Spearman	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8	LP9	LP10
lsa:- dim:- sco:count sw:yes	0.886	0.714	0.771	0.886	0.714	0.086	0.486	0.486	0.543	0.486
lsa:- dim:- sco:count sw:-	0.943	0.371	0.943	0.086	0.657	0.886	0.600	0.257	0.600	0.200
lsa:- dim:- sco:tfidf sw:yes	0.714	0.771	0.771	0.886	0.886	0.257	0.714	0.600	0.829	0.429
lsa:- dim:- sco:tfidf sw:-	0.771	0.657	0.771	0.029	0.657	0.943	0.714	0.257	0.714	0.543
lsa:- dim:- sco:zeroone sw:yes	0.714	0.714	0.543	0.886	0.771	0.257	0.771	0.486	0.257	0.429
lsa:- dim:- sco:zeroone sw:-	1.000	0.771	0.771	0.371	0.943	1.000	0.943	-0.086	0.829	0.714
lsa:yes dim:100 sco:count sw:yes	0.943	0.200	0.771	0.886	0.657	0.257	0.486	0.600	0.486	0.086
lsa:yes dim:100 sco:count sw:-	1.000	0.371	0.943	0.086	0.600	0.943	0.657	0.029	0.600	0.200
lsa:yes dim:100 sco:tfidf sw:yes	0.829	0.200	0.943	0.886	0.600	0.257	0.486	0.600	0.543	0.371
lsa:yes dim:100 sco:tfidf sw:-	1.000	0.600	0.943	0.029	0.657	0.943	0.600	0.371	0.714	0.200
lsa:yes dim:100 sco:zeroone sw:yes	0.829	0.314	0.829	0.829	0.829	0.600	0.543	0.257	0.600	0.257
lsa:yes dim:100 sco:zeroone sw:-	0.886	0.714	0.771	0.257	1.000	0.829	0.371	0.086	0.829	0.600
lsa:yes dim:300 sco:count sw:yes	0.771	0.200	0.886	0.886	0.657	0.257	0.257	0.657	0.486	0.486
lsa:yes dim:300 sco:count sw:-	1.000	0.371	0.943	0.086	0.600	0.943	0.600	0.029	0.600	0.200
lsa:yes dim:300 sco:tfidf sw:yes	0.943	0.314	0.943	0.829	0.657	0.600	0.257	0.657	0.486	0.486
lsa:yes dim:300 sco:tfidf sw:-	1.000	0.657	0.943	0.086	0.600	0.943	0.714	0.257	0.886	0.371
lsa:yes dim:300 sco:zeroone sw:yes	0.714	0.543	0.886	1.000	0.657	0.886	0.714	0.657	0.829	0.600
lsa:yes dim:300 sco:zeroone sw:-	0.943	0.886	0.714	0.371	1.000	0.943	0.657	-0.086	0.829	0.943

Results

I. Zero-one, sw, dim:300

II. TF-IDF, sw

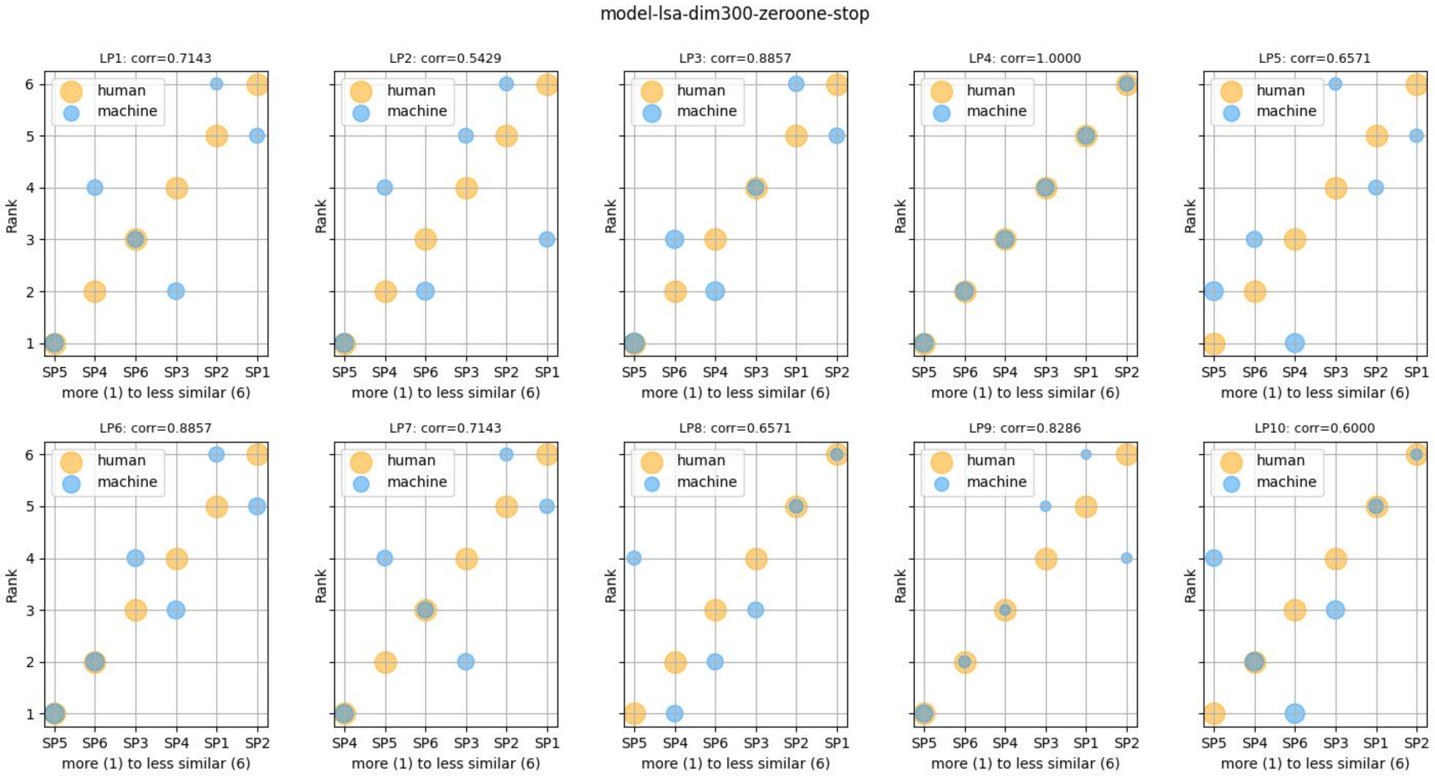
III. TF-IDF, sw, dim:300

X. Zero-one, sw

Green: highest values on avg
Red: lowest values (0.3 up) on avg
Yellow: low values (0-0.3) on stdev

	Spearman		Kendall		O P Measure	
Model	AVG	ST DEV	AVG	ST DEV	AVG	ST DEV
lsa:- dim:- sco:count sw:yes	0.606	0.242	0.467	0.208	0.733	0.104
lsa:- dim:- sco:count sw:-	0.554	0.315	0.480	0.284	0.740	0.142
lsa:- dim:- sco:tfidf sw:yes (II)	0.686	0.204	0.533	0.201	0.767	0.101
lsa:- dim:- sco:tfidf sw:-	0.606	0.270	0.507	0.260	0.753	0.130
lsa:- dim:- sco:zeroone sw:yes (X)	0.583	0.222	0.427	0.189	0.713	0.095
lsa:- dim:- sco:zeroone sw:-	0.726	0.341	0.627	0.360	0.813	0.180
lsa:yes dim:100 sco:count sw:yes	0.537	0.291	0.413	0.296	0.707	0.148
lsa:yes dim:100 sco:count sw:-	0.543	0.360	0.480	0.358	0.740	0.179
lsa:yes dim:100 sco:tfidf sw:yes	0.571	0.256	0.427	0.267	0.713	0.134
lsa:yes dim:100 sco:tfidf sw:-	0.606	0.325	0.533	0.328	0.767	0.164
lsa:yes dim:100 sco:zeroone sw:yes	0.589	0.242	0.480	0.222	0.740	0.111
lsa:yes dim:100 sco:zeroone sw:-	0.634	0.300	0.520	0.275	0.760	0.138
lsa:yes dim:300 sco:count sw:yes	0.554	0.258	0.440	0.242	0.720	0.121
lsa:yes dim:300 sco:count sw:-	0.537	0.359	0.467	0.356	0.733	0.178
lsa:yes dim:300 sco:tfidf sw:yes (III)	0.617	0.240	0.507	0.244	0.753	0.122
lsa:yes dim:300 sco:tfidf sw:-	0.646	0.317	0.573	0.307	0.787	0.153
lsa:yes dim:300 sco:zeroone sw:yes (I)	0.749	0.146	0.613	0.183	0.807	0.091
lsa:yes dim:300 sco:zeroone sw:-	0.720	0.340	0.613	0.352	0.807	0.176

Results - Ranking Comparison. One model.



[Models on GitHub]

Graph of LPx and rankings of SP(i,x), given by humans and the machine.

Circle area = similarity value.

Conclusion and Further work

- Stopwords improve results. Each model has its own threshold
- Best results: *LSA(dim:300), zero-one and stopwords*

1. **Human Ranking**: Larger population to obtain better human's perception
2. **Test Data set**: Clear differentiation between similar sentences
3. **Corpus**: Vocabulary and knowledge
   - Use a corpus that gives the system enough background on the topic to be tested
4. **Generating our models**:
   - No syntactic information: "Computer Bag" or "Bag the computer", the same. Try n-grams
5. **Storage:** Truncated LSA-models
   - Try Stemming and other preprocessing techniques to reduce matrix size

Thank you!

Results - Similarities - LP2

1 0.5 0

LP2	For the majority of women, breast cancer is not a death sentence. In the U.S., fewer women have been dying from breast cancer since the 1990s. Women whose breast cancer has not spread to other organs in their body like the lungs, liver, or brain, have over a 90% chance of being alive in 5 years. How long each person diagnosed with breast cancer will live and whether she will die from the disease depends on many things, including how much the cancer has spread throughout the body, the woman's overall health, and more.
SP(1,2)	Cancer are malignant cells that grow into more cells to form a tumor which needs to be treated to not become lethal.
SP(2,2)	Systemic therapy is for the whole body and can target and kill cancer in other parts of the body.
SP(3,2)	Depending on the advancement of breast cancer as well of other factors, the patient might survive.
SP(4,2)	The perception of breast cancer is changing everyday. Women are not afraid because they know that if the diagnosis is early, they will have the opportunity to cure it depending on the comorbidities.
SP(5,2)	For most of the women population who have been diagnosed with breast cancer, it is no longer a life threatening situation. Those whom their cancer has not spread to other organs like lungs, liver or brain have a 90% possibility of living the next five years or so.
SP(6,2)	Breast cancer mortality depends on many factors, like health, lifestyle, and spread to other organs, but cancer is less deadly than 20-30 years ago.

model		LP2					
		SP5	SP6	SP4	SP1	SP2	SP3
lsa:- dim:- sco:count sw:yes	1	0.754	0.749	0.579	0.502	0.449	0.299
lsa:- dim:- sco:count sw:-	2	0.995	0.993	0.991	0.982	0.995	0.986
lsa:- dim:- sco:tfidf sw:yes	3	0.678	0.664	0.524	0.393	0.401	0.366
lsa:- dim:- sco:tfidf sw:-	4	0.986	0.969	0.972	0.953	0.982	0.962
lsa:- dim:- sco:zeroone sw:yes	5	0.854	0.837	0.722	0.553	0.535	0.392
lsa:- dim:- sco:zeroone sw:-	6	0.991	0.977	0.975	0.968	0.973	0.968
lsa:yes dim:100 sco:count sw:yes	7	0.711	0.747	0.491	0.635	0.552	0.244
lsa:yes dim:100 sco:count sw:-	8	0.994	0.991	0.988	0.979	0.995	0.986
lsa:yes dim:100 sco:tfidf sw:yes	9	0.729	0.760	0.520	0.635	0.584	0.299
lsa:yes dim:100 sco:tfidf sw:-	10	0.988	0.979	0.976	0.963	0.987	0.970
lsa:yes dim:100 sco:zeroone sw:yes	11	0.899	0.897	0.768	0.846	0.795	0.720
lsa:yes dim:100 sco:zeroone sw:-	12	0.996	0.988	0.982	0.980	0.980	0.979
lsa:yes dim:300 sco:count sw:yes	13	0.681	0.714	0.483	0.608	0.562	0.288
lsa:yes dim:300 sco:count sw:-	14	0.994	0.991	0.988	0.978	0.995	0.985
lsa:yes dim:300 sco:tfidf sw:yes	15	0.673	0.631	0.452	0.580	0.548	0.332
lsa:yes dim:300 sco:tfidf sw:-	16	0.987	0.974	0.974	0.959	0.985	0.968
lsa:yes dim:300 sco:zeroone sw:yes	17	0.862	0.829	0.707	0.713	0.647	0.689
lsa:yes dim:300 sco:zeroone sw:-	18	0.993	0.982	0.979	0.973	0.977	0.973

Results - Similarities - LP4

	1	0.5	0

LP4	So far herbal products have not been shown to cure cancer. However, there are some herbs that may help patients deal with the side effects of cancer treatments. We recommend that you talk to your doctor before taking any vitamins and herbal products as they might have an effect on your cancer treatments.
SP(1,4)	Even though your attitude might not be a direct cause of cancer, it is helpful to stay positive and social.
SP(2,4)	There are many variables to determine whether a woman will die from breast cancer within the next five years.
SP(3,4)	Herbal products and vitamins are not linked to any treatment for cancer, however, it can minimize the secondary effects of proper treatment.
SP(4,4)	People are accustomed to consuming herbal products but there is no evidence of cure when they use them. However they could help with side effects of traditional cancer treatments.
SP(5,4)	So far, there is no proof that herbal products heal cancer, but there are some that can help with the treatment side effects. Nevertheless any herbal or vitamin products that you might want to try should be advised by your treating physician.
SP(6,4)	Herbal products may help with some cancer side effects, but you should discuss them with your doctor before starting to take any vitamins or supplements.

model		LP9					
		SP5	SP4	SP3	SP2	SP6	SP1
lsa:- dim:- sco:count sw:yes	1	0.694	0.520	0.477	0.460	0.409	0.388
lsa:- dim:- sco:count sw:-	2	0.996	0.996	0.996	0.984	0.997	0.994
lsa:- dim:- sco:tfidf sw:yes	3	0.548	0.385	0.356	0.301	0.413	0.394
lsa:- dim:- sco:tfidf sw:-	4	0.985	0.980	0.985	0.955	0.990	0.977
lsa:- dim:- sco:zeroone sw:yes	5	0.665	0.468	0.414	0.452	0.288	0.302
lsa:- dim:- sco:zeroone sw:-	6	0.982	0.973	0.973	0.968	0.986	0.962
lsa:yes dim:100 sco:count sw:yes	7	0.920	0.806	0.850	0.688	0.410	0.379
lsa:yes dim:100 sco:count sw:-	8	0.995	0.995	0.995	0.981	0.997	0.993
lsa:yes dim:100 sco:tfidf sw:yes	9	0.916	0.828	0.816	0.649	0.455	0.432
lsa:yes dim:100 sco:tfidf sw:-	10	0.988	0.986	0.989	0.963	0.993	0.981
lsa:yes dim:100 sco:zeroone sw:yes	11	0.940	0.749	0.868	0.760	0.843	0.678
lsa:yes dim:100 sco:zeroone sw:-	12	0.992	0.985	0.989	0.985	0.994	0.975
lsa:yes dim:300 sco:count sw:yes	13	0.891	0.707	0.833	0.650	0.441	0.393
lsa:yes dim:300 sco:count sw:-	14	0.995	0.995	0.995	0.981	0.997	0.993
lsa:yes dim:300 sco:tfidf sw:yes	15	0.822	0.667	0.734	0.604	0.478	0.436
lsa:yes dim:300 sco:tfidf sw:-	16	0.987	0.982	0.986	0.956	0.992	0.980
lsa:yes dim:300 sco:zeroone sw:yes	17	0.798	0.489	0.484	0.489	0.553	0.457
lsa:yes dim:300 sco:zeroone sw:-	18	0.986	0.976	0.979	0.973	0.990	0.965

Results - Similarities - LP5

1 0.5 0

LP5	Breast cancer is a type of cancer that starts in the breast. Cancer starts when cells begin to grow out of control. It's important to understand that most breast lumps are benign and not cancer (malignant). Non-cancerous breast tumors are abnormal growths, but they do not spread outside of the breast. They are not life threatening, but some types of benign breast lumps can increase a woman's risk of getting breast cancer. Any breast lump or change needs to be checked by a healthcare professional to determine if it is benign or malignant (cancer) and if it might affect your future cancer risk.
SP(1,5)	Herbs might have beneficial effects on cancer treatments.
SP(2,5)	Getting good exercise, having a positive attitude, and letting your emotions out are good for those who have had cancer.
SP(3,5)	The growth of malignant cells on the breast is called breast cancer, however, not all are malignant but can indicate a predisposition.
SP(4,5)	Many types of breast tumor are benign. It's important to do a regular and frequent medical check up to figure out if it's malignant.
SP(5,5)	Cancer begins when body cells begin to grow out of control. This can happen in any part of the body, but when this happens in the breast tissue this is called breast cancer. Any breast lump or masses should be checked by a specialist to determine if it's malignant.
SP(6,5)	A lump or change in the breast may be cancer or it may not be -- those non cancerous lumps are not damaging to your health, but only a doctor can tell what kind of lump it is.

model		LP5					
		SP5	SP4	SP2	SP6	SP3	SP1
lsa:- dim:- sco:count sw:yes	1	0.900	0.897	0.555	0.534	0.517	0.279
lsa:- dim:- sco:count sw:-	2	0.995	0.997	0.990	0.997	0.996	0.945
lsa:- dim:- sco:tfidf sw:yes	3	0.675	0.663	0.448	0.482	0.371	0.271
lsa:- dim:- sco:tfidf sw:-	4	0.987	0.988	0.972	0.990	0.988	0.837
lsa:- dim:- sco:zeroone sw:yes	5	0.809	0.810	0.666	0.697	0.546	0.325
lsa:- dim:- sco:zeroone sw:-	6	0.993	0.990	0.982	0.993	0.990	0.938
lsa:yes dim:100 sco:count sw:yes	7	0.967	0.958	0.645	0.515	0.708	0.282
lsa:yes dim:100 sco:count sw:-	8	0.993	0.995	0.986	0.997	0.995	0.928
lsa:yes dim:100 sco:tfidf sw:yes	9	0.949	0.927	0.677	0.567	0.666	0.344
lsa:yes dim:100 sco:tfidf sw:-	10	0.988	0.991	0.977	0.993	0.991	0.847
lsa:yes dim:100 sco:zeroone sw:yes	11	0.955	0.918	0.785	0.847	0.892	0.701
lsa:yes dim:100 sco:zeroone sw:-	12	0.997	0.996	0.988	0.997	0.994	0.985
lsa:yes dim:300 sco:count sw:yes	13	0.950	0.941	0.564	0.504	0.651	0.297
lsa:yes dim:300 sco:count sw:-	14	0.993	0.995	0.986	0.997	0.995	0.930
lsa:yes dim:300 sco:tfidf sw:yes	15	0.904	0.865	0.583	0.552	0.634	0.386
lsa:yes dim:300 sco:tfidf sw:-	16	0.988	0.990	0.974	0.993	0.990	0.837
lsa:yes dim:300 sco:zeroone sw:yes	17	0.875	0.876	0.689	0.745	0.582	0.620
lsa:yes dim:300 sco:zeroone sw:-	18	0.995	0.992	0.984	0.995	0.991	0.947

Results - Similarities - LP6

1	0.5	0

LP6	For most types of cancer, doctors use staging information to help plan treatment and to predict a person's outlook (prognosis). Although each person's situation is different, cancers with the same stage tend to have similar outlooks and are often treated the same way. The cancer stage is also a way for doctors to describe the extent of the cancer when they talk with each other about a person's cancer.
SP(1,6)	Even though not all new lumps or masses are malignant, they should be checked by a doctor to determine its risks.
SP(2,6)	Before taking herbal products, that may help some deal with cancer side effects, patients should consult with their doctor.
SP(3,6)	The advancement of cancer is categorized and described in stages, which also works as a guide for treatment and prediction.
SP(4,6)	Doctors usually use stages for cancer so they can predict the evolution and can choose the treatment. Actually they can describe the extension of the tumor through staging.
SP(5,6)	When doctors are dealing with cancer patients, Oncologists use staging information as a reference to determine the severity of the patient, and their treatment as well. Even though each patient has their own characteristics, there are similar symptoms during each disease's stage.
SP(6,6)	Each person with cancer is unique, but cancer stages allow doctors to group similar patients together when they discuss treatment options and outcomes.

model		LP6					
		SP5	SP4	SP2	SP1	SP6	SP3
lsa:- dim:- sco:count sw:yes	1	0.888	0.615	0.602	0.579	0.526	0.384
lsa:- dim:- sco:count sw:-	2	0.996	0.986	0.966	0.986	0.994	0.994
lsa:- dim:- sco:tfidf sw:yes	3	0.720	0.499	0.463	0.356	0.462	0.335
lsa:- dim:- sco:tfidf sw:-	4	0.987	0.969	0.905	0.962	0.977	0.985
lsa:- dim:- sco:zeroone sw:yes	5	0.861	0.716	0.709	0.645	0.706	0.497
lsa:- dim:- sco:zeroone sw:-	6	0.988	0.979	0.947	0.977	0.981	0.979
lsa:yes dim:100 sco:count sw:yes	7	0.945	0.690	0.722	0.569	0.691	0.521
lsa:yes dim:100 sco:count sw:-	8	0.996	0.986	0.959	0.984	0.992	0.993
lsa:yes dim:100 sco:tfidf sw:yes	9	0.900	0.768	0.775	0.535	0.739	0.569
lsa:yes dim:100 sco:tfidf sw:-	10	0.992	0.975	0.944	0.969	0.987	0.988
lsa:yes dim:100 sco:zeroone sw:yes	11	0.939	0.886	0.896	0.752	0.928	0.880
lsa:yes dim:100 sco:zeroone sw:-	12	0.996	0.988	0.969	0.986	0.993	0.985
lsa:yes dim:300 sco:count sw:yes	13	0.907	0.620	0.647	0.548	0.644	0.479
lsa:yes dim:300 sco:count sw:-	14	0.996	0.986	0.958	0.984	0.992	0.993
lsa:yes dim:300 sco:tfidf sw:yes	15	0.803	0.655	0.664	0.435	0.689	0.542
lsa:yes dim:300 sco:tfidf sw:-	16	0.990	0.973	0.914	0.963	0.984	0.987
lsa:yes dim:300 sco:zeroone sw:yes	17	0.900	0.831	0.785	0.710	0.845	0.785
lsa:yes dim:300 sco:zeroone sw:-	18	0.991	0.985	0.956	0.979	0.988	0.982

Results - Similarities - LP9

1 0.5 0

LP9	A mastectomy is a type of surgery to remove the entire breast. There are several types of mastectomies. Some mastectomies remove all of the breast tissue, including the nipple and areola and the other types of mastectomies also remove some of the nearby lymph nodes
SP(1,9)	A lumpectomy is surgery to remove the cancer tumor from the breast.
SP(2,9)	It is difficult to determine by a new lumps shape or hardness whether it is cancer or not.
SP(3,9)	There are different types of mastectomy, depending on how much of the breast they remove, up to the complete removal.
SP(4,9)	Mastectomy is a way to get rid of breast cancer and it would be as radical depending on how far the cancer has spread. The surgeons may take out tissue as long is compromised with the disease.
SP(5,9)	The mastectomy is a surgery that removes all the breast tissue. There are different types of mastectomy depending on the area removed. The first one focuses on all breast tissues including the nipple and areola, while others also add nearby lymph nodes to the list.
SP(6,9)	Mastectomies are surgeries that remove some or all of the breast, and may also remove adjacent lymph nodes.

model		LP9					
		SP5	SP4	SP3	SP2	SP6	SP1
lsa:- dim:- sco:count sw:yes	1	0.694	0.520	0.477	0.460	0.409	0.388
lsa:- dim:- sco:count sw:-	2	0.996	0.996	0.996	0.984	0.997	0.994
lsa:- dim:- sco:tfidf sw:yes	3	0.548	0.385	0.356	0.301	0.413	0.394
lsa:- dim:- sco:tfidf sw:-	4	0.985	0.980	0.985	0.955	0.990	0.977
lsa:- dim:- sco:zeroone sw:yes	5	0.665	0.468	0.414	0.452	0.288	0.302
lsa:- dim:- sco:zeroone sw:-	6	0.982	0.973	0.973	0.968	0.986	0.962
lsa:yes dim:100 sco:count sw:yes	7	0.920	0.806	0.850	0.688	0.410	0.379
lsa:yes dim:100 sco:count sw:-	8	0.995	0.995	0.995	0.981	0.997	0.993
lsa:yes dim:100 sco:tfidf sw:yes	9	0.916	0.828	0.816	0.649	0.455	0.432
lsa:yes dim:100 sco:tfidf sw:-	10	0.988	0.986	0.989	0.963	0.993	0.981
lsa:yes dim:100 sco:zeroone sw:yes	11	0.940	0.749	0.868	0.760	0.843	0.678
lsa:yes dim:100 sco:zeroone sw:-	12	0.992	0.985	0.989	0.985	0.994	0.975
lsa:yes dim:300 sco:count sw:yes	13	0.891	0.707	0.833	0.650	0.441	0.393
lsa:yes dim:300 sco:count sw:-	14	0.995	0.995	0.995	0.981	0.997	0.993
lsa:yes dim:300 sco:tfidf sw:yes	15	0.822	0.667	0.734	0.604	0.478	0.436
lsa:yes dim:300 sco:tfidf sw:-	16	0.987	0.982	0.986	0.956	0.992	0.980
lsa:yes dim:300 sco:zeroone sw:yes	17	0.798	0.489	0.484	0.489	0.553	0.457
lsa:yes dim:300 sco:zeroone sw:-	18	0.986	0.976	0.979	0.973	0.990	0.965

Results - Kendall Tau Correlation

Kendall Tau	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8	LP9	LP10
lsa:- dim:- sco:count sw:yes	0.733	0.467	0.600	0.733	0.600	0.067	0.333	0.333	0.467	0.333
lsa:- dim:- sco:count sw:-	0.867	0.333	0.867	0.067	0.600	0.733	0.467	0.200	0.467	0.200
lsa:- dim:- sco:tfidf sw:yes	0.467	0.600	0.600	0.733	0.733	0.200	0.600	0.467	0.733	0.200
lsa:- dim:- sco:tfidf sw:-	0.600	0.600	0.600	-0.067	0.600	0.867	0.600	0.200	0.600	0.467
lsa:- dim:- sco:zeroone sw:yes	0.467	0.467	0.467	0.733	0.600	0.200	0.600	0.333	0.200	0.200
lsa:- dim:- sco:zeroone sw:-	1.000	0.600	0.600	0.333	0.867	1.000	0.867	-0.200	0.600	0.600
lsa:yes dim:100 sco:count sw:yes	0.867	0.067	0.600	0.733	0.600	0.200	0.333	0.467	0.333	-0.067
lsa:yes dim:100 sco:count sw:-	1.000	0.333	0.867	0.067	0.467	0.867	0.600	-0.067	0.467	0.200
lsa:yes dim:100 sco:tfidf sw:yes	0.600	0.067	0.867	0.733	0.467	0.200	0.333	0.467	0.467	0.067
lsa:yes dim:100 sco:tfidf sw:-	1.000	0.467	0.867	-0.067	0.600	0.867	0.467	0.333	0.600	0.200
lsa:yes dim:100 sco:zeroone sw:yes	0.600	0.200	0.733	0.733	0.733	0.467	0.467	0.200	0.467	0.200
lsa:yes dim:100 sco:zeroone sw:-	0.733	0.467	0.600	0.200	1.000	0.733	0.333	0.067	0.600	0.467
lsa:yes dim:300 sco:count sw:yes	0.600	0.067	0.733	0.733	0.600	0.200	0.200	0.600	0.333	0.333
lsa:yes dim:300 sco:count sw:-	1.000	0.333	0.867	0.067	0.467	0.867	0.467	-0.067	0.467	0.200
lsa:yes dim:300 sco:tfidf sw:yes	0.867	0.200	0.867	0.600	0.600	0.467	0.200	0.600	0.333	0.333
lsa:yes dim:300 sco:tfidf sw:-	1.000	0.600	0.867	0.067	0.467	0.867	0.600	0.200	0.733	0.333
lsa:yes dim:300 sco:zeroone sw:yes	0.467	0.467	0.733	1.000	0.467	0.733	0.467	0.600	0.733	0.467
lsa:yes dim:300 sco:zeroone sw:-	0.867	0.733	0.467	0.333	1.000	0.867	0.600	-0.200	0.600	0.867

Results - Order Preservation Measure

Order Preservation Measure	LP1	LP2	LP3	LP4	LP5	LP6	LP7	LP8	LP9	LP10
lsa:- dim:- sco:count sw:yes	0.867	0.733	0.800	0.867	0.800	0.533	0.667	0.667	0.733	0.667
lsa:- dim:- sco:count sw:-	0.933	0.667	0.933	0.533	0.800	0.867	0.733	0.600	0.733	0.600
lsa:- dim:- sco:tfidf sw:yes	0.733	0.800	0.800	0.867	0.867	0.600	0.800	0.733	0.867	0.600
lsa:- dim:- sco:tfidf sw:-	0.800	0.800	0.800	0.467	0.800	0.933	0.800	0.600	0.800	0.733
lsa:- dim:- sco:zeroone sw:yes	0.733	0.733	0.733	0.867	0.800	0.600	0.800	0.667	0.600	0.600
lsa:- dim:- sco:zeroone sw:-	1.000	0.800	0.800	0.667	0.933	1.000	0.933	0.400	0.800	0.800
lsa:yes dim:100 sco:count sw:yes	0.933	0.533	0.800	0.867	0.800	0.600	0.667	0.733	0.667	0.467
lsa:yes dim:100 sco:count sw:-	1.000	0.667	0.933	0.533	0.733	0.933	0.800	0.467	0.733	0.600
lsa:yes dim:100 sco:tfidf sw:yes	0.800	0.533	0.933	0.867	0.733	0.600	0.667	0.733	0.733	0.533
lsa:yes dim:100 sco:tfidf sw:-	1.000	0.733	0.933	0.467	0.800	0.933	0.733	0.667	0.800	0.600
lsa:yes dim:100 sco:zeroone sw:yes	0.800	0.600	0.867	0.867	0.867	0.733	0.733	0.600	0.733	0.600
lsa:yes dim:100 sco:zeroone sw:-	0.867	0.733	0.800	0.600	1.000	0.867	0.667	0.533	0.800	0.733
lsa:yes dim:300 sco:count sw:yes	0.800	0.533	0.867	0.867	0.800	0.600	0.600	0.800	0.667	0.667
lsa:yes dim:300 sco:count sw:-	1.000	0.667	0.933	0.533	0.733	0.933	0.733	0.467	0.733	0.600
lsa:yes dim:300 sco:tfidf sw:yes	0.933	0.600	0.933	0.800	0.800	0.733	0.600	0.800	0.667	0.667
lsa:yes dim:300 sco:tfidf sw:-	1.000	0.800	0.933	0.533	0.733	0.933	0.800	0.600	0.867	0.667
lsa:yes dim:300 sco:zeroone sw:yes	0.733	0.733	0.867	1.000	0.733	0.867	0.733	0.800	0.867	0.733
lsa:yes dim:300 sco:zeroone sw:-	0.933	0.867	0.733	0.667	1.000	0.933	0.800	0.400	0.800	0.933